(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 February 2002 (07.02.2002)

PCT

(10) International Publication Number
WO 02/11123 A2

(51) International Patent Classification[7]: G10L 17/00

(21) International Application Number: PCT/EP01/08709

(22) International Filing Date: 26 July 2001 (26.07.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/222,023 31 July 2000 (31.07.2000) US
09/839,476 20 April 2001 (20.04.2001) US

(71) Applicant *(for all designated States except US)*: **SHAZAM ENTERTAINMENT LIMITED** [GB/GB]; 189 Wardour Street, Suite 22, London W1F 8ZD (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **WANG, Avery, Li-Chun** [US/US]; 2915 Ross Road, Palo Alto, CA 94303 (US). **SMITH, Julius, O. III** [US/US]; 4360 Miller Avenue, Palo Alto, CA 94308 (US).

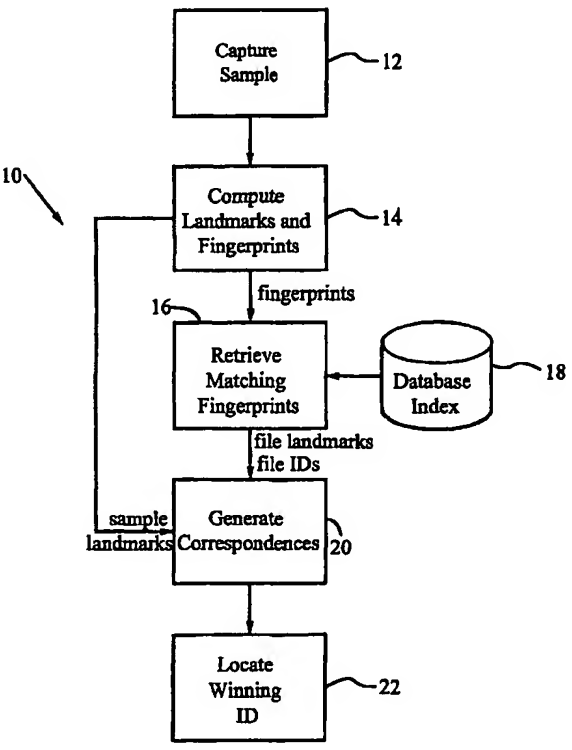
(74) Agent: **BOYCE, Conor**; F. R. Kelly & Co., 27 Clyde Road, Ballsbridge, Dublin 4 (IE).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF,

[Continued on next page]

(54) Title: SYSTEM AND METHODS FOR RECOGNIZING SOUND AND MUSIC SIGNALS IN HIGH NOISE AND DISTORTION



(57) Abstract: A method for recognizing an audio sample locates an audio file that most closely matches the audio sample from a database indexing a large set of original recordings. Each indexed audio file is represented in the database index by a set of landmark timepoints and associated fingerprints. Landmarks occur at reproductible locations within the file, while fingerprints represent features of the signal at or near the landmark timepoints. To perform recognition, landmarks and fingerprints are computed for the unknown sample and used to retrieve matching fingerprints from the database. For each file containing matching fingerprints, the landmarks are compared with landmarks of the sample at which the same fingerprints were computed. If a large number of corresponding landmarks are linearly related, i.e., if equivalent fingerprints of the sample and retrieved file have the same time evolution, then the file is identified with the sample. The method can be used for any type of sound or music, and is particularly effective for audio signals subject to linear and nonlinear distortion such as background noise, compression artifacts, or transmission dropouts. The sample can be identified in a time proportional to the logarithm of the number of entries in the database; given sufficient computational power, recognition can be performed in nearly real time as the sound is being sampled.

WO 02/11123 A2

CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii)) for all designations*

— *of inventorship (Rule 4.17(iv)) for US only*

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

SYSTEM AND METHODS FOR RECOGNIZING SOUND AND MUSIC SIGNALS IN HIGH NOISE AND DISTORTION

FIELD OF THE INVENTION

This invention relates generally to content-based information retrieval. More particularly, it relates to recognition of an audio signal, including sound or music, that is highly distorted or contains a high level of noise.

BACKGROUND ART

There is a growing need for automatic recognition of music or other audio signals generated from a variety of sources. For example, owners of copyrighted works or advertisers are interested in obtaining data on the frequency of broadcast of their material. Music tracking services provide playlists of major radio stations in large markets. Consumers would like to identify songs or advertising broadcast on the radio, so that they can purchase new and interesting music or other products and services. Any sort of continual or on-demand sound recognition is inefficient and labor intensive when performed by humans. An automated method of recognizing music or sound would thus provide significant benefit to consumers, artists, and a variety of industries. As the music distribution paradigm shifts from store purchases to downloading via the Internet, it is quite feasible to link directly computer-implemented music recognition with Internet purchasing and other Internet-based services.

Traditionally, recognition of songs played on the radio has been performed by matching radio stations and times at which songs were played with playlists provided either by the radio stations or from third party sources. This method is inherently limited to only radio stations for which information is available. Other methods rely on embedding inaudible codes within broadcast signals. The embedded signals are decoded at the receiver to extract identifying information about the broadcast signal. The disadvantage of this method is that special decoding devices are required to identify signals, and only those songs with embedded codes can be identified.

Any large-scale audio recognition requires some sort of content-based audio retrieval, in which an unidentified broadcast signal is compared with a database of known signals to identify similar or identical database signals. Note that content-based audio retrieval is different from existing audio retrieval by web search engines, in which only the metadata text surrounding or associated with audio files is searched. Also note that while speech recognition is useful for converting voiced signals into text that can then be indexed and

searched using well-known techniques, it is not applicable to the large majority of audio signals that contain music and sounds. In some ways, audio information retrieval is analogous to text-based information retrieval provided by search engines. In other ways, however, audio recognition is not analogous: audio signals lack easily identifiable entities such as words that provide identifiers for searching and indexing. As such, current audio retrieval schemes index audio signals by computed perceptual characteristics that represent various qualities or features of the signal.

Content-based audio retrieval is typically performed by analyzing a query signal to obtain a number of representative characteristics, and then applying a similarity measure to the derived characteristics to locate database files that are most similar to the query signal. The similarity of received objects is necessarily a reflection of the perceptual characteristics selected. A number of content-based retrieval methods are available in the art. For example, U.S. Patent No. 5,210,820, issued to Kenyon, discloses a signal recognition method in which received signals are processed and sampled to obtain signal values at each sampling point. Statistical moments of the sampled values are then computed to generate a feature vector that can be compared with identifiers of stored signals to retrieve similar signals. U.S. Patent Nos. 4,450,531 and 4,843,562, both issued to Kenyon et al., disclose similar broadcast information classification methods in which cross-correlations are computed between unidentified signals and stored reference signals.

A system for retrieving audio documents by acoustic similarity is disclosed in J. T. Foote, "Content-Based Retrieval of Music and Audio," in C.-C. J. Kuo et al., editor, *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, volume 3229, pages 138-147, 1997. Feature vectors are calculated by parameterizing each audio file into mel-scaled cepstral coefficients, and a quantization tree is grown from the parameterization data. To perform a query, an unknown signal is parameterized to obtain feature vectors that are then sorted into leaf nodes of the tree. A histogram is collected for each leaf node, thereby generating an N-dimensional vector representing the unknown signal. The distance between two such vectors is indicative of the similarity between two sound files. In this method, the supervised quantization scheme learns distinguishing audio features, while ignoring unimportant variations, based on classes into which the training data are assigned by a human. Depending upon the classification system, different acoustic features are chosen to be important. Thus this method is more suited for finding similarities between songs and sorting music into classes than it is to recognizing music.

A method for content-based analysis, storage, retrieval, and segmentation of audio information is disclosed in U.S. Patent No. 5,918,223, issued to Blum et al. In this method, a number of acoustical features, such as loudness, bass, pitch, brightness, bandwidth, and Mel-frequency cepstral coefficients, are measured at periodic intervals of each file. Statistical measurements of the features are taken and combined to form a feature vector. Audio data files within a database are retrieved based on the similarity of their feature vectors to the feature vector of an unidentified file.

A key problem of all of the above prior art audio recognition methods is that they tend to fail when the signals to be recognized are subject to linear and nonlinear distortion caused by, for example, background noise, transmission errors and dropouts, interference, band-limited filtering, quantization, time-warping, and voice-quality digital compression. In prior art methods, when a distorted sound sample is processed to obtain acoustical features, only a fraction of the features derived for the original recording are found. The resulting feature vector is therefore not very similar to the feature vector of the original recording, and it is unlikely that correct recognition can be performed. There remains a need for a sound recognition system that performs well under conditions of high noise and distortion.

Another problem with prior art methods is that they are computationally intensive and do not scale well. Real-time recognition is thus not possible using prior art methods with large databases. In such systems, it is unfeasible to have a database of more than a few hundred or thousand recordings. Search time in prior art methods tends to grow linearly with the size of the database, making scaling to millions of sounds recordings economically unfeasible. The methods of Kenyon also require large banks of specialized digital signal processing hardware.

Existing commercial methods often have strict requirements for the input sample to be able to perform recognition. For example, they require the entire song or at least 30 seconds of the song to be sampled or require the song to be sampled from the beginning. They also have difficulty recognizing multiple songs mixed together in a single stream. All of these disadvantages make prior art methods unfeasible for use in many practical applications.

OBJECTS AND ADVANTAGES

Accordingly, it is a primary object of the present invention to provide a method for recognizing an audio signal subject to a high level of noise and distortion.

It is a further object of the invention to provide a recognition method that can be performed in real time based on only a few seconds of the signal to be identified.

It is another object of the invention to provide a recognition method than can recognize sounds based on samples from almost anywhere within the sound, not just at the beginning.

It is an additional object of the invention to provide a recognition method that does not require sound samples to be coded or correlated with particular radio stations or playlists.

It is a further object of the invention to provide a recognition method that can recognize each of multiple sound recordings mixed together in a single stream.

It is another object of the invention to provide a sound recognition system in which the unknown sound can be provided to the system from any environment by virtually any known method.

SUMMARY

These objects and advantages are attained by a method for recognizing a media sample, such as an audio sample, given a database index of a large number of known media files. The database index contains fingerprints representing features at particular locations of the indexed media files. The unknown media sample is identified with a media file in the database (the winning media file) whose relative locations of fingerprints most closely match the relative locations of fingerprints of the sample. In the case of audio files, the time evolution of fingerprints of the winning file matches the time evolution of fingerprints in the sample.

The method is preferably implemented in a distributed computer system and contains the following steps: determining a set of fingerprints at particular locations of the sample; locating matching fingerprints in the database index; generating correspondences between locations in the sample and locations in the file having equivalent fingerprints; and identifying media files for which a significant number of the correspondences are substantially linearly related. The file having the largest number of linearly related correspondences is deemed the winning media file. One method of identifying files with a large number of correspondences is to perform the equivalent of scanning for a diagonal line in the scatter plot generated from the pairs of correspondences. In one embodiment,

identifying the media files with a large number of linear correspondences involves searching only a first subset of the media files. Files in the first subset have a higher probability of being identified than files that are not in the first subset. The probability of identification is preferably based on empirical frequency or recency measures of previous identifications, along with *a priori* projections of identification frequency. If no media files are identified in the first subset, then the second subset, containing the remaining files, is searched. Alternatively, the files can be ranked by probability and searched in order of the ranking. The search is terminated when a file is located.

Preferably, the particular locations within the sample are reproducibly computed in dependence on the sample. Such reproducibly computable locations are called "landmarks." Fingerprints are preferably numerical values. In one embodiment, each fingerprint represents a number of features of the media sample at each location, or offset slightly from the location.

The method is particularly useful for recognizing audio samples, in which case the particular locations are timepoints within the audio sample. These timepoints occur at, for example, local maxima of spectral Lp norms of the audio sample. Fingerprints can be computed by any analysis of the audio sample, and are preferably invariant to time stretching of the sample. Examples of fingerprints include spectral slice fingerprints, multi-slice fingerprints, LPC coefficients, cepstral coefficients, and frequency components of spectrogram peaks.

The present invention also provides a system for implementing the above method, containing a landmarking object for computing the particular locations, a fingerprinting object for computing the fingerprints, a database index containing the file locations and fingerprints for the media files, and an analysis object. The analysis object implements the method by locating matching fingerprints in the database index, generating correspondences, and analyzing the correspondences to select the winning media file.

Also provided is a program storage device accessible by a computer, tangibly embodying a program of instructions executable by the computer to perform method steps for the above method.

Additionally, the invention provides a method for creating an index of a number of audio files in a database, containing the following steps: computing a set of fingerprints at particular locations of each file; and storing the fingerprints, locations, and identifiers of

the files in a memory. A corresponding fingerprint, location, and identifier is associated in the memory to form a triplet. Preferably, the locations, which can be timepoints within the audio file, are computed in dependence on the file and are reproducible. For example, the timepoints can occur at local maxima of spectral Lp norms of the audio file. In some cases, each fingerprint, which is preferably a numerical value, represents a number of features of the file near the particular location. Fingerprints can be computed from any analysis or digital signal processing of the audio file. Examples of fingerprints include spectral slice fingerprints, multi-slice fingerprints, LPC coefficients, cepstral coefficients, frequency components of spectrogram peaks, and linked spectrogram peaks.

Finally, the invention provides methods for identifying audio samples incorporating time-stretch invariant fingerprints and various hierarchical searching.

BRIEF DESCRIPTION OF THE FIGURES

Fig. **1** is a flow diagram of a method of the invention for recognizing a sound sample.

Fig. **2** is a block diagram of an exemplary distributed computer system for implementing the method of Fig. 1.

Fig. **3** is a flow diagram of a method for constructing a database index of sound files used in the method of Fig. 1.

Fig. **4** schematically illustrates landmarks and fingerprints computed for a sound sample.

Fig. **5** is a graph of L4 norms for a sound sample, illustrating the selection of landmarks.

Fig. **6** is a flow diagram of an alternative embodiment for constructing a database index of sound files used in the method of Fig. 1.

Figs. **7A-7C** show a spectrogram with salient points and linked salient points indicated.

Figs. **8A-8C** illustrate index sets, an index list, and a master index list of the method of Fig. 3.

Figs. **9A-9C** illustrate an index list, candidate list, and scatter list of the method of Fig. 1.

Figs. **10A-10B** are scatter plots illustrating correct identification and lack of identification, respectively, of an unknown sound sample.

DETAILED DESCRIPTION

The present invention provides a method for recognizing an exogenous media sample given a database containing a large number of known media files. It also provides a method for generating a database index that allows efficient searching using the recognition method of the invention. While the following discussion refers primarily to audio data, it is to be understood that the method of the present invention can be applied to any type of media samples and media files, including, but not limited to, text, audio,

video, image, and any multimedia combinations of individual media types. In the case of audio, the present invention is particularly useful for recognizing samples that contain high levels of linear and nonlinear distortion caused by, for example, background noise, transmission errors and dropouts, interference, band-limited filtering, quantization, time-warping, and voice-quality digital compression. As will be apparent from the description below, the invention works under such conditions because it can correctly recognize a distorted signal even if only a small fraction of the computed characteristics survive the distortion. Any type of audio, including sound, voice, music, or combinations of types, can be recognized by the present invention. Example audio samples include recorded music, radio broadcast programs, and advertisements.

As used herein, an exogenous media sample is a segment of media data of any size obtained from a variety of sources as described below. In order for recognition to be performed, the sample must be a rendition of part of a media file indexed in a database used by the present invention. The indexed media file can be thought of as an original recording, and the sample as a distorted and/or abridged version or rendition of the original recording. Typically, the sample corresponds to only a small portion of the indexed file. For example, recognition can be performed on a ten-second segment of a five-minute song indexed in the database. Although the term "file" is used to describe the indexed entity, the entity can be in any format for which the necessary values (described below) can be obtained. Furthermore, there is no need to store or have access to the file after the values are obtained.

A block diagram conceptually illustrating the overall steps of a method **10** of the present invention is shown in **Fig. 1.** Individual steps are described in more detail below. The method identifies a winning media file, a media file whose relative locations of characteristic fingerprints most closely match the relative locations of the same fingerprints of the exogenous sample. After an exogenous sample is captured in step **12,** landmarks and fingerprints are computed in step **14.** Landmarks occur at particular locations, e.g., timepoints, within the sample. The location within the sample of the landmarks is preferably determined by the sample itself, i.e., is dependent upon sample qualities, and is reproducible. That is, the same landmarks are computed for the same signal each time the process is repeated. For each landmark, a fingerprint characterizing one or more features of the sample at or near the landmark is obtained. The nearness of a feature to a landmark is defined by the fingerprinting method used. In some cases, a feature is considered near a landmark if it clearly corresponds to the landmark and not to a previous or subsequent landmark. In other cases, features correspond to multiple adjacent

landmarks. For example, text fingerprints can be word strings, audio fingerprints can be spectral components, and image fingerprints can be pixel RGB values. Two general embodiments of step **14** are described below, one in which landmarks and fingerprints are computed sequentially, and one in which they are computed simultaneously.

In step **16,** the sample fingerprints are used to retrieve sets of matching fingerprints stored in a database index **18,** in which the matching fingerprints are associated with landmarks and identifiers of a set of media files. The set of retrieved file identifiers and landmark values are then used to generate correspondence pairs (step **20**) containing sample landmarks (computed in step **14**) and retrieved file landmarks at which the same fingerprints were computed. The resulting correspondence pairs are then sorted by song identifier, generating sets of correspondences between sample landmarks and file landmarks for each applicable file. Each set is scanned for alignment between the file landmarks and sample landmarks. That is, linear correspondences in the pairs of landmarks are identified, and the set is scored according to the number of pairs that are linearly related. A linear correspondence occurs when a large number of corresponding sample locations and file locations can be described with substantially the same linear equation, within an allowed tolerance. For example, if the slopes of a number of equations describing a set of correspondence pairs vary by ±5%, then the entire set of correspondences is considered to be linearly related. Of course, any suitable tolerance can be selected. The identifier of the set with the highest score, i.e., with the largest number of linearly related correspondences, is the winning file identifier, which is located and returned in step **22.**

As described further below, recognition can be performed with a time component proportional to the logarithm of the number of entries in the database. Recognition can be performed in essentially real time, even with a very large database. That is, a sample can be recognized as it is being obtained, with a small time lag. The method can identify a sound based on segments of 5-10 seconds and even as low 1-3 seconds. In a preferred embodiment, the landmarking and fingerprinting analysis, step **14,** is carried out in real time as the sample is being captured in step **12.** Database queries (step **16**) are carried out as sample fingerprints become available, and the correspondence results are accumulated and periodically scanned for linear correspondences. Thus all of the method steps occur simultaneously, and not in the sequential linear fashion suggested in Fig. 1. Note that the method is in part analogous to a text search engine: a user submits a query sample, and a matching file indexed in the sound database is returned.

The method is typically implemented as software running on a computer system, with individual steps most efficiently implemented as independent software modules. Thus a system implementing the present invention can be considered to consist of a landmarking and fingerprinting object, an indexed database, and an analysis object for searching the database index, computing correspondences, and identifying the winning file. In the case of sequential landmarking and fingerprinting, the landmarking and fingerprinting object can be considered to be distinct landmarking and fingerprinting objects. Computer instruction code for the different objects is stored in a memory of one or more computers and executed by one or more computer processors. In one embodiment, the code objects are clustered together in a single computer system, such as an Intel-based personal computer or other workstation. In a preferred embodiment, the method is implemented by a networked cluster of central processing units (CPUs), in which different software objects are executed by different processors in order to distribute the computational load. Alternatively, each CPU can have a copy of all software objects, allowing for a homogeneous network of identically configured elements. In this latter configuration, each CPU has a subset of the database index and is responsible for searching its own subset of media files.

Although the invention is not limited to any particular hardware system, an example of a preferred embodiment of a distributed computer system **30** is illustrated schematically in **Fig. 2.** System **30** contains a cluster of Linux-based processors **32a-32f** connected by a multiprocessing bus architecture **34** or a networking protocol such as the Beowulf cluster computing protocol, or a mixture of the two. In such an arrangement, the database index is preferably stored in random access memory (RAM) on at least one node **32a** in the cluster, ensuring that fingerprint searching occurs very rapidly. The computational nodes corresponding to the other objects, such as landmarking nodes **32c** and **32f,** fingerprinting nodes **32b** and **32e,** and alignment scanning node **32d,** do not require as much bulk RAM as does node or nodes **32a** supporting the database index. The number of computational nodes assigned to each object may thus be scaled according to need so that no single object becomes a bottleneck. The computational network is therefore highly parallelizable and can additionally process multiple simultaneous signal recognition queries that are distributed among available computational resources. Note that this makes possible applications in which large numbers of users can request recognition and receive results in near real time.

In an alternative embodiment, certain of the functional objects are more tightly coupled together, while remaining less tightly coupled to other objects. For example, the

landmarking and fingerprinting object can reside in a physically separate location from the rest of the computational objects. One example of this is a tight association of the landmarking and fingerprinting objects with the signal capturing process. In this arrangement, the landmarking and fingerprinting object can be incorporated as additional hardware or software embedded in, for example, a mobile phone, Wireless Application Protocol (WAP) browser, personal digital assistant (PDA), or other remote terminal, such as the client end of an audio search engine. In an Internet-based audio search service, such as a content identification service, the landmarking and fingerprinting object can be incorporated into the client browser application as a linked set of software instructions or a software plug-in module such as a Microsoft dynamic link library (DLL). In these embodiments, the combined signal capture, landmarking, and fingerprinting object constitutes the client end of the service. The client end sends a feature-extracted summary of the captured signal sample containing landmark and fingerprint pairs to the server end, which performs the recognition. Sending this feature-extracted summary to the server, instead of the raw captured signal, is advantageous because the amount of data is greatly reduced, often by a factor of 500 or more. Such information can be sent in real time over a low-bandwidth side channel along with or instead of, e.g., an audio stream transmitted to the server. This enables performing the invention over public communications networks, which offer relatively small-sized bandwidths to each user.

The method will now be described in detail with reference to audio samples and audio files indexed in a sound database. The method consists of two broad components, sound database index construction and sample recognition.

Database index construction

Before sound recognition can be performed, a searchable sound database index must be constructed. As used herein, a database is any indexed collection of data, and is not limited to commercially available databases. In the database index, related elements of data are associated with one another, and individual elements can be used to retrieve associated data. The sound database index contains an index set for each file or recording in the selected collection or library of recordings, which may include speech, music, advertisements, sonar signatures, or other sounds. Each recording also has a unique identifier, `sound_ID`. The sound database itself does not necessarily store the audio files for each recording, but the `sound_IDs` can be used to retrieve the audio files from elsewhere. The sound database index is expected to be very large, containing indices for millions or even billions of files. New recordings are preferably added incrementally to the database index.

A block diagram of a preferred method **40** for constructing the searchable sound database index according to a first embodiment is shown in **Fig. 3.** In this embodiment, landmarks are first computed, and then fingerprints are computed at or near the landmarks. As will be apparent to one of average skill in the art, alternative methods may be devised for constructing the database index. In particular, many of the steps listed below are optional, but serve to generate a database index that is more efficiently searched. While searching efficiency is important for real-time sound recognition from large databases, small databases can be searched relatively quickly even if they have not been sorted optimally.

To index the sound database, each recording in the collection is subjected to a landmarking and fingerprinting analysis that generates an index set for each audio file. **Fig. 4** schematically illustrates a segment of a sound recording for which landmarks (LM) and fingerprints (FP) have been computed. Landmarks occur at specific timepoints of the sound and have values in time units offset from the beginning of the file, while fingerprints characterize the sound at or near a particular landmark. Thus, in this embodiment, each landmark for a particular file is unique, while the same fingerprint can occur numerous times within a single file or multiple files.

In step **42,** each sound recording is landmarked using methods to find distinctive and reproducible locations within the sound recording. A preferred landmarking algorithm is able to mark the same timepoints within a sound recording despite the presence of noise and other linear and nonlinear distortion. Some landmarking methods are conceptually independent of the fingerprinting process described below, but can be chosen to optimize performance of the latter. Landmarking results in a list of timepoints $\{\texttt{landmark}_k\}$ within the sound recording at which fingerprints are subsequently calculated. A good landmarking scheme marks about 5-10 landmarks per second of sound recording; of course, landmarking density depends on the amount of activity within the sound recording.

A variety of techniques are possible for computing landmarks, all of which are within the scope of the present invention. The specific technical processes used to implement the landmarking schemes of the invention are known in the art and will not be discussed in detail. A simple landmarking technique, known as Power Norm, is to calculate the instantaneous power at every possible timepoint in the recording and to select local maxima. One way of doing this is to calculate the envelope by rectifying and filtering the waveform directly. Another way is to calculate the Hilbert transform (quadrature) of the

signal and use the sum of the magnitudes squared of the Hilbert transform and the original signal.

The Power Norm method of landmarking is good at finding transients in the sound signal. The Power Norm is actually a special case of the more general Spectral Lp Norm in which p=2. The general Spectral Lp Norm is calculated at each time along the sound signal by calculating a short-time spectrum, for example via a Hanning-windowed Fast Fourier Transform (FFT). A preferred embodiment uses a samping rate of 8000Hz, an FFT frame size of 1024 samples, and a stride of 64 samples for each time slice. The Lp norm for each time slice is then calculated as the sum of the p^{th} power of the absolute values of the spectral components, optionally followed by taking the p^{th} root. As before, the landmarks are chosen as the local maxima of the resulting values over time. An example of the Spectral Lp Norm method is shown in **Fig. 5,** a graph of the L4 norm as a function of time for a particular sound signal. Dashed lines at local maxima indicate the location of the chosen landmarks.

When $p=\infty$, the $L\infty$ norm is effectively the maximum norm. That is, the value of the norm is the absolute value of the largest spectral component in the spectral slice. This norm results in robust landmarks and good overall recognition performance, and is preferred for tonal music.

Alternatively, "multi-slice" spectral landmarks can be calculated by taking the sum of p^{th} powers of absolute values of spectral components over multiple timeslices at fixed or variable offsets from each other, instead of a single slice. Finding the local maxima of this extended sum allows optimization of placement of the multi-slice fingerprints, described below.

Once the landmarks have been computed, a fingerprint is computed at each landmark timepoint in the recording in step **44.** The fingerprint is generally a value or set of values that summarizes a set of features in the recording at or near the timepoint. In a currently preferred embodiment, each fingerprint is a single numerical value that is a hashed function of multiple features. Possible types of fingerprints include spectral slice fingerprints, multi-slice fingerprints, LPC coefficients, and cepstral coefficients. Of course, any type of fingerprint that characterizes the signal or features of the signal near a landmark is within the scope of the present invention. Fingerprints can be computed by any type of digital signal processing or frequency analysis of the signal.